Universiteit Leiden
Governance and Global Affairs

Mapping Digital Governance and Artificial Intelligence in European Parliamentary Debates

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation at the Workshop on The Politics of AI
Institut Barcelona d'Estudis Internacionals (IBEI), Barcelona, October 2–3, 2025

Introduction

Research Overview

Corpus scale ∼7.5M speeches from 26 European countries analysed using advanced ML techniques.

Proof-of-concept. Four-country study (2015-2022): Bulgaria, the Czech Republic, Estonia, and Spain, representing different digital profiles.

Methodology. 118 models benchmarked (115 LLMs + 3 ft-BERTs) against 1,000 hand-coded bills to ensure classification reliability.

Artwork by DALL·E 3 model

The Digital Governance Challenge

Artwork by DALL·E 3 model

Promise of AI. Governments rapidly integrate digital tools for efficiency gains, improved service delivery, and data-driven decision-making.

Emerging risks. AI raises concerns about transparency, accountability, bias amplification, and discriminatory policy outcomes.

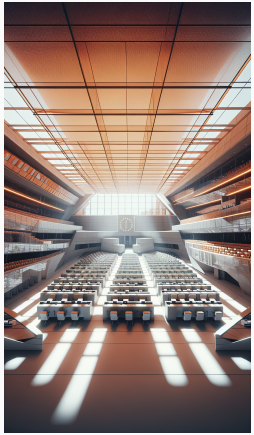
Research gap. We know comparatively little about how MPs talk about digital governance over time across countries, and whether some events (random shocks) reconfigure the parliamentary agenda.

Focus on Salience, not Stance

Measure attention. Establish when and how much parliaments talk about digital technologies.

Understand patterns. Legislative time and floor space are scarce, thus, changes in attention should precede policy action.

Future analysis. Stance classification (risks vs opportunities) is reserved for the next iteration of the paper.



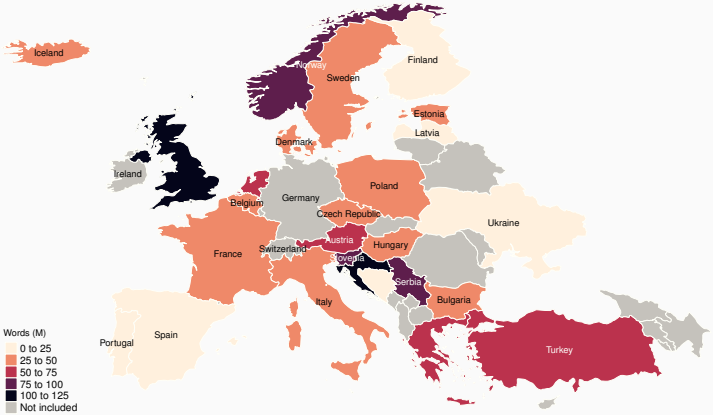
Artwork by DALL·E 3 model

Data and Methods

Parliamentary Debates Data

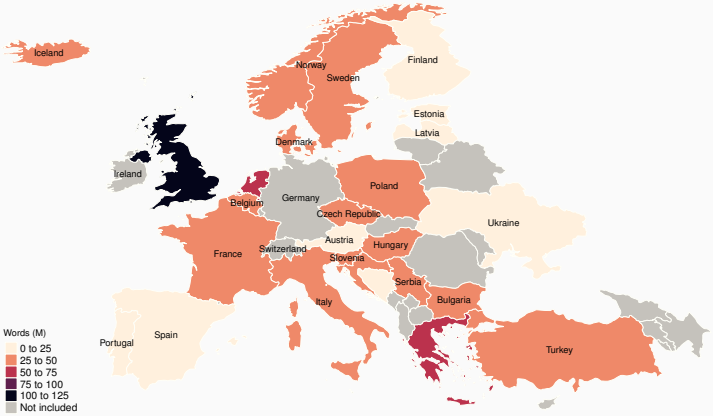
We use ParlaMint, which offers data on parliamentary debates of 26 European countries and three Spanish autonomous regions from about 2015 to mid-2022 (some countries even from the 1990s). The dataset offers not only the original speeches but also corpora linguistically annotated in the original languages and machine-translated versions of the texts.

ParlaMint Corpora ∼1.1B words



This implies that **∼7.5M utterances** were delivered in Parliament in the case of unicameral systems or some of the relevant chambers in bicameral systems.

Subsetting Corpora ∼912M words



Austria, Bosnia and Herzegovina, Norway, and Serbia have had coverage since the 1990s. Therefore, to balance better corpora for cross-national comparison, we sliced the data and **focused on the 2015-2022 period**.

Coverage and Machine Translation Accuracy

- The filtered machine-translated corpora cover ∼**6M** of utterances and ∼**912M words**.
- The upper chamber is not included in some of bicameral systems, namely Austria, Belgium, Czech Republic and France. In Italy, the *Camera* is not included.
- Low machine-translated quality (BLEU $\leq$ 40) for Slovenian, Norwegian and Bosnian.
- In general, the quality of machine translation is high and fluent (50 $\leq$ BLEU $<$ 60), with several languages standing out with quality that is probably better than an average human's (BLEU $\geq$ 60). E.g., Danish, Italian.

* BLEU: Bilingual Evaluation Understudy score.

Case Selection Strategy

Bulgaria: Laggard case (BLEU 61.6)

Low capability (35.5% digital skills) and low e-government use (36.5%)

Czech Republic: Capability-led (BLEU 62.5)

Top-tier digital skills (69.1%) but weaker infrastructure (49.1% broadband)

Estonia: Platform-led (BLEU 61.8)

World-class egovernment uptake (94.4%) despite modest broadband coverage (48.4%)

Spain: All-round frontrunner (BLEU 46.4)

Robust broadband (95.7%), e-government use (82.76%), solid capability

* European Commission's Digital Economy and Society Index (DESI) indicators.

LLMs Zero-Shot Classification

115 LLMs

3 fine-tuned BERTs

1,000 hand-coded bills
as ground-truth labels for evaluation

SOTA closed-source LLMs

GPT-5, Grok 4, o4-mini, o3-mini, GPT-4.1, Claude 4 Opus, Gemini 2.5, etc.

SOTA open-source LLMs

GPT-OSS (20 and 120B), Qwen 3 (235B), Llama 4 Maverick (400B) and Scout (107B), Mistral Medium 3.1, DeepSeek-V3.1 (671B), etc.

Fine-tuned BERTs

ft-ModernBERT, CAP Babel Machine, ft-XLM-RoBERTa

Difference-in-Differences Design

› **Stacked DiD.** Monthly frequency exploiting common-timed shocks across countries. Three-month post-event windows analysed.

› **Uncertainty propagation.** Central, lower-bound, and upper-bound attention k-th series constructed from the two best classifiers.

$$Y^k_{c[t]} = \alpha_c + \gamma_t + \sum_{e=1}^{E} \delta_e^{k[Spain]} (Spain \times Post_{e[t]}) + \varepsilon_{c[t]} \tag{1}$$

$$Y^k_{c[t]} = \alpha_c + \gamma_t + \sum_{e=1}^{E} \delta_e^{k[Bulgaria]} (Bulgaria \times Post_{e[t]}) + \varepsilon_{c[t]} \tag{2}$$

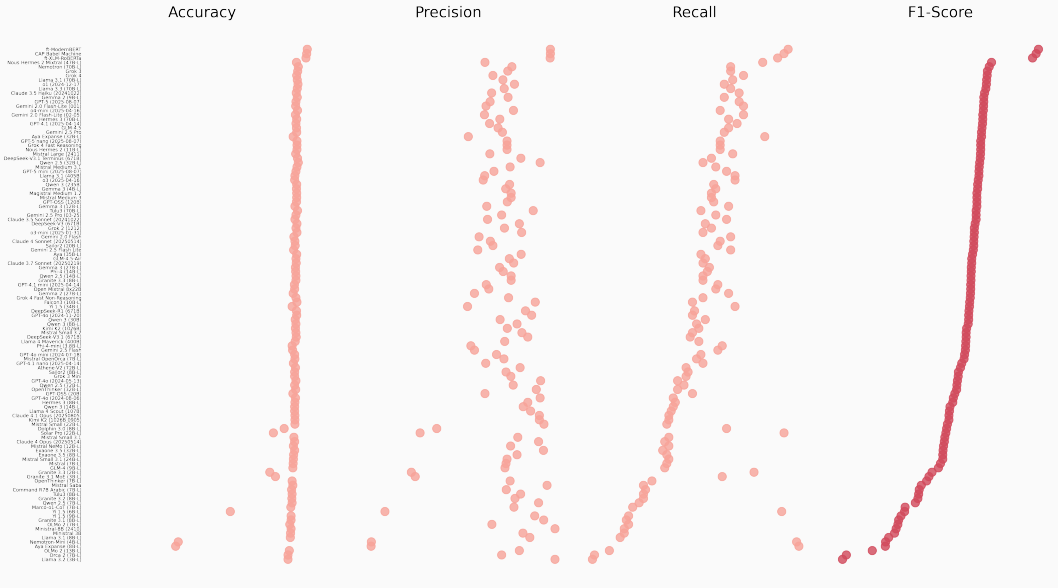
Key Technological Shocks (2015-2022)

- **March-May 2018.** Cambridge Analytica revelations and GDPR enforcement reshape the data protection landscape.
- **February 2020.** European Commission's White Paper on AI signals forthcoming legislation.
- **June 2020.** GPT-3 public release demonstrates advanced language model capabilities.
- **April 2021.** EU AI Act proposal introduces risk-based regulatory framework.

GPT-3.5/ChatGTP

We avoided reliance on the late-2022 diffusion of GPT-3.5/ChatGPT, which falls at the end of our period, but may be incorporated into future iterations of the paper.

Preliminary Results


Accuracy
Precision
Recall
F1-Score
0.0
0.5
1.0

Benchmark Highlights

0.96

ft-ModernBERT

Maximum F1-score
even after overlap
adjustement 0.86

0.95

CAP Babel Machine

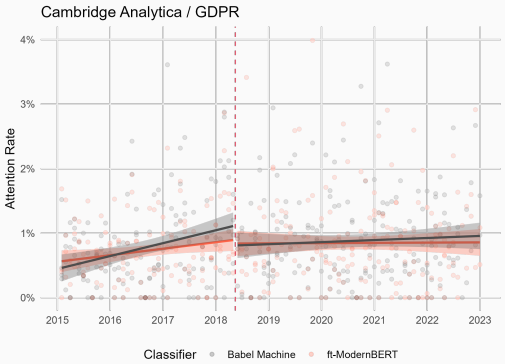
Strong second
even after overlap
adjustement 0.82

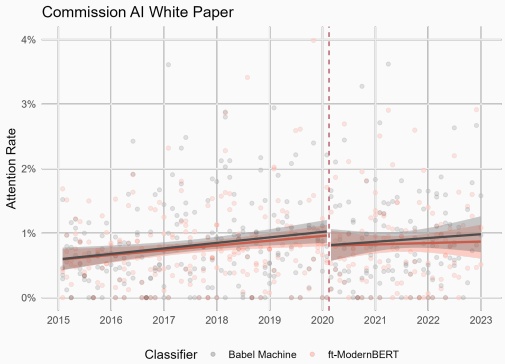
0.74

Best LLMs

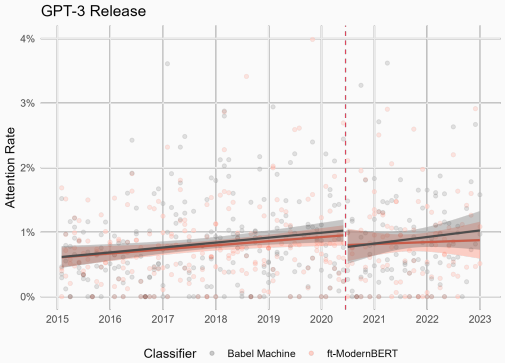
Best performers cluster
around F1 0.69-0.74,
behind fine-tuned encoders

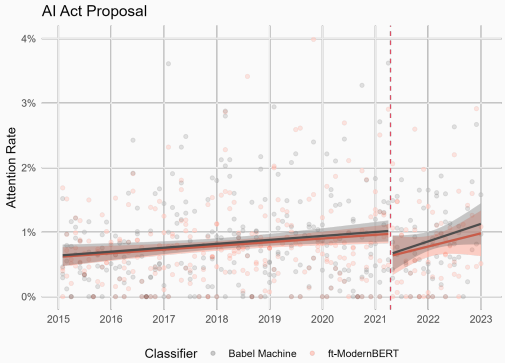
Low Baseline Attention and No Clear Pattern





Low Baseline Attention and Delayed Response





Effect of Punctuations on Parliamentary Attention

	Attention Rate					
	Lower	Central	Upper	Lower	Central	Upper
Frontrunner × CA/GDPR	1.181*	1.342	1.503			
	(0.659)	(0.869)	(1.077)			
Frontrunner × AI White Paper	−0.075	−0.163	−0.250			
	(0.241)	(0.265)	(0.302)			
Frontrunner × GPT-3 release	0.118	0.182	0.247			
	(0.436)	(0.445)	(0.453)			
Frontrunner × AI Act proposal	−0.490***	−0.442***	−0.393*			
	(0.094)	(0.166)	(0.216)			
Laggard × CA/GDPR				−0.492	−0.591	−0.690
				(0.629)	(0.659)	(0.683)
Laggard × AI White Paper				−0.079	0.126	0.332
				(0.213)	(0.249)	(0.288)
Laggard × GPT-3 release				0.660***	0.749***	0.837**
				(0.108)	(0.241)	(0.348)
Laggard × AI Act proposal				−0.361	−0.428**	−0.494***
				(0.289)	(0.206)	(0.120)
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
N	337	337	337	337	337	337
R^2	0.403	0.412	0.412	0.388	0.397	0.399
Adj. R^2	0.143	0.155	0.155	0.121	0.134	0.138

Effect of Punctuations on Parliamentary Attention

	Attention Rate					
	Lower	Central	Upper	Lower	Central	Upper
Frontrunner × CA/GDPR	1.181*	1.342	1.503			
	(0.659)	(0.869)	(1.077)			
Frontrunner × AI White Paper	−0.075	−0.163	−0.250			
	(0.241)	(0.265)	(0.302)			
Frontrunner × GPT-3 release	0.118	0.182	0.247			
	(0.436)	(0.445)	(0.453)			
Frontrunner × AI Act proposal	−0.490***	−0.442***	−0.393*			
	(0.094)	(0.166)	(0.216)			
Laggard × CA/GDPR		**−83%** vs baseline		−0.492	−0.591	−0.690
				(0.629)	(0.659)	(0.683)
Laggard × AI White Paper				−0.079	0.126	0.332
				(0.213)	(0.249)	(0.288)
Laggard × GPT-3 release			**+82%** vs baseline	0.660***	0.749***	0.837**
				(0.108)	(0.241)	(0.348)
Laggard × AI Act proposal				−0.361	−0.428**	−0.494***
				(0.289)	(0.206)	(0.120)
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
N	337	337	337	337	337	337
R^2	0.403	0.412	0.412	0.388	0.397	0.399
Adj. R^2	0.143	0.155	0.155	0.121	0.134	0.138

Takeaways and Next Steps

Takeaways

› **Comprehensive benchmarking.** It provides a transparent foundation for classifier selection in parliamentary text analysis.

› **Uncertainty propagation.** Approach using multiple classifiers to construct confidence bounds for downstream DiD.

➜ Incorporate human-in-the-loop validation.

› **Overlap adjustment.** Scoring adjustment addresses potential data leakage in fine-tuned models.

› **Empirical insights.** Low baseline attention amplifies relative effects of technological shocks.

Limitations and Future Directions

› **Translation effects.** Machine-translated corpora may smooth country-specific linguistic nuances despite high BLEU scores.

➜ Re-estimate on original language corpora.

› **Scale expansion.** Update to the latest ParlaMint release, incorporating all 26 countries and extended time periods.

➜ Incorporate late-2022 GPT-3.5/ChatGPT and later SOTA models diffusion effects.

› **Stance classification.** Move from salience to stance classification (e.g., risks vs opportunities).

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub